KU LEUVEN

Centrum voor Taal en Onderwijs

Creating meaning out of vagueness?

Using the CEFR as the foundation for an academic rating scale

Bart Deygers & Koen Van Gorp - CLE, KU Leuven

Contents

I Rating scales and rater variability

Rating scales are tools

- Some are based on intuition
  (and may be too vague)
- Some are based on performance data
  (and may be too detailed)
- Some are based on both

(Knoch, 2009; Galaczi et al., 2011; Fulcher, 2012)

Rating scales are tools, used by raters

Raters differ

- in the focus of their attention when rating a performance
- in their interpretation of the same rating scale
- in their compliance to its criteria,
- in their severity when assigning scores based on those criteria

(Weigle, 2002; Lumley, 2005; Eckes, 2008)

Reducing rater variability

Option 1: Train your raters

But rater training does not always reduce rater variability

Option 2: Employ experienced raters

But experience does not always guarantee reliability

Option 3: Involve your raters

"The best scales will be those built by the raters themselves. So rather than giving them a scale, I'd try to develop a scale in meetings with the raters"

Bernard Spolsky, Ltest-L, 12 May 2013

(Weigle, 2002; Derwing et al., 2004; Elder et al., 2005; Lumley, 2005; Eckes, 2008; Barkaoui, 2010; Isaacs & Thomson, 2013)

II The CEFR as the basis for a rating scale

Frameworks, scales & reliability

- Generalizing descriptions of language are abstractions by definition

(Lumley, 2002; Fulcher, 2004; Dubeau, 2006; Alderson, 2007; Hulstijn, 2007; Little, 2007; Eckes, 2008; Roberts, Donoghue & Laughlin, 2010; Galaczi et al., 2011; Fulcher, 2012; North, 2014; Zeidler, 2014)

Frameworks, scales & reliability

- Generalizing descriptions of language are abstractions by definition
- Frameworks are too unspecified to be applied directly as rating scales

(Lumley, 2002; Fulcher, 2004; Dubeau, 2006; Alderson, 2007; Hulstijn, 2007; Little, 2007; Eckes, 2008; Roberts, Donoghue & Laughlin, 2010; Galaczi et al., 2011; Fulcher, 2012; North, 2014; Zeidler, 2014)

Frameworks, scales & reliability

- Generalizing descriptions of language are abstractions by definition
- Frameworks are too underspecified to be applied directly as rating scales
- Frameworks do not automatically mean the same to different people

(Lumley, 2002; Fulcher, 2004; Dubeau, 2006; Alderson, 2007; Hulstijn, 2007; Little, 2007; Eckes, 2008; Roberts, Donoghue & Laughlin, 2010; Galaczi et al., 2011; Fulcher, 2012; North, 2014; Zeidler, 2014)

The case of the CEFR

- Limited basis in SLA and empirical research
- Overfocused on production
- Too generic nature and impressionistic
- Too many inconsistencies

It's too vague, really

(Weir, 2005; Alderson, 2007; Little, 2007; Roberts, Donoghue & Laughlin, 2010; Fulcher, 2012)

Eeny, meeny, miny, moe?

ALTE Study (CEFR SIG):

Population:

10 CEFR experts

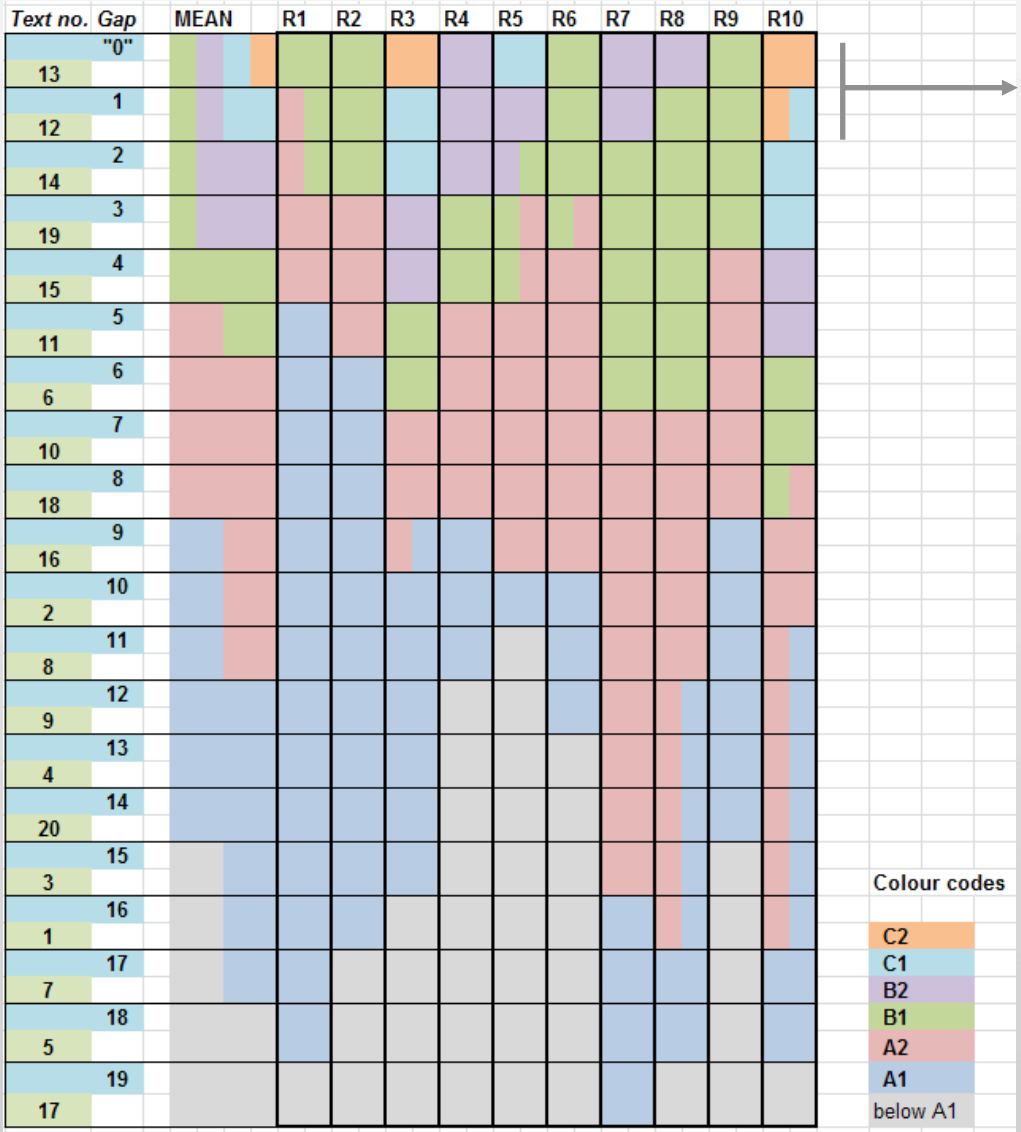
Samples:

20 short written English texts
Identical task
(L1 = Finnish)

Method:

Rank order and assign level

Eeny, meeny, miny, moe?



A1 -> C2

No consensus for any sample

Zeidler & Deygers, 2012

The CEFR as a rating scale

The CEFR can't be applied directly

And it needs to be reworked to fit the context of use

(Galaczi et al., 2011; North, 2014)

III The study

Context of the study

Certificate of Dutch as a Foreign Language (CNaVT):

- Worldwide, 3500 candidates/year
- 5 tests (5 contexts, 4 levels)
- 2014: New tests & analytic rating scales

  Funding organization demand: CEFR-based rating scales

Today's focus: Dutch for Academic Purposes, B2

Research Questions

RQ1: Is it possible to develop a reliable DAP rating scale that is based on CEFR descriptors?

RQ2: Does an empirical co-construction process with novice raters help to stimulate a shared interpretation of CEFR-based criteria?

(What happened before)

Iterative rating scale co-construction (2010, 2011, 2012, 2013)

(Weir, 1990; Deygers & Van Gorp, 2013)

(What happened before)

Iterative rating scale co-construction (2010, 2011, 2012, 2013)

4-band scale

Target level +1	C1
Target level	B2
Target level -1	B1
Target level -2	A2

(Weir, 1990; Deygers & Van Gorp, 2013)

(What happened before)

Iterative rating scale co-construction (2010, 2011, 2012, 2013)

4-band scale

Target level +1	C1
Target level	B2
Target level -1	B1
Target level -2	A2

CEFR-based descriptors, supplemented with:

- exemplars
- concrete insertions
- criterion definitions

(Deygers & Van Gorp, 2013)

Method & procedure

6 trained novice raters

↓

Rater training & rating scale co-construction

Method & procedure

6 trained novice raters

200 samples, stratified for level, L1, country

Writing: summary & argumentation

Spoken: presentation & argumentation

Method & procedure

6 trained novice raters

200 samples, stratified for level, L1, country etc…

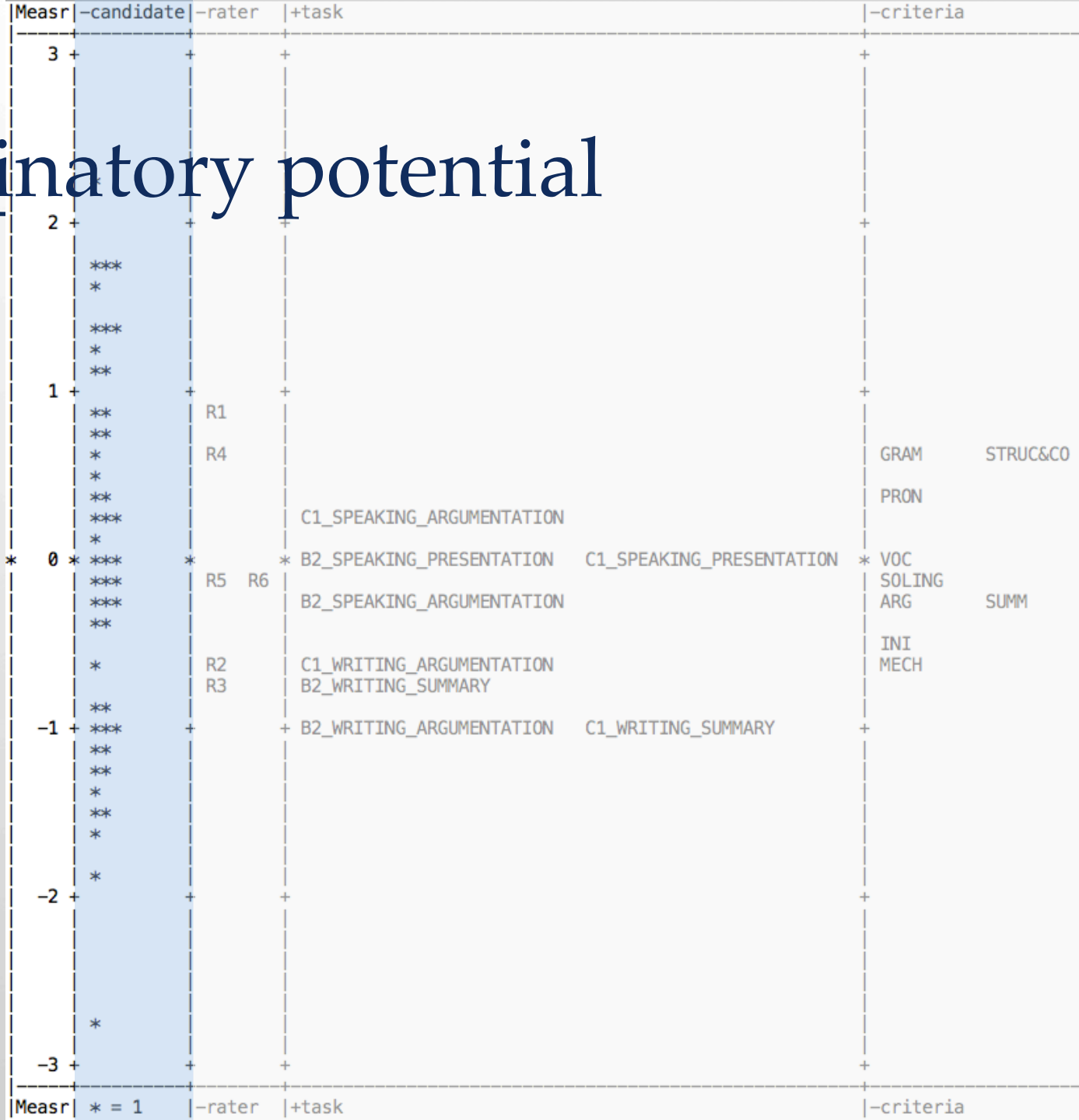
Mixed-method data collection & analysis

Quantitative:	Descriptive data
	Discriminatory potential
	Rater uniformity
	Rating variability
	Principal component analysis
Qualitative:	Focus group

(Brenner & Kliebst, 1996; Sim & Wright, 2005 Vieira, Kaymak & Sousa, 2010; Linacre, 2012)

RQ1 Is it possible to develop a reliable DAP rating scale that is based on CEFR descriptors?

Discriminatory potential



Rater uniformity & rating variability

- Rater separation: differing severity (-.70 - .90)
- High inter-rater agreement (Kw = .8)

- Acceptable rater variability (InfitMnSq = .93 – 1.12)

Criteria

- Levels of difficulty (-.65 – .69)
- Criteria fit the Rasch model (InfitMnSq = .74 – 1.51)

Criterion	Measure	S.E.	Infit MnSq
Structure & cohesion	.69	.08	1.11
Grammar	.66	.08	.85
Pronunciation	.37	.11	1.07
Vocabulary	-.01	.07	.86
Sociolinguistics	-.12	.15	.81
Summarizing	-.19	.15	1.04
Argumentation	-.27	.10	1.09
Initiative	-.48	.11	.74
Mechanics	-.65	.10	1.51

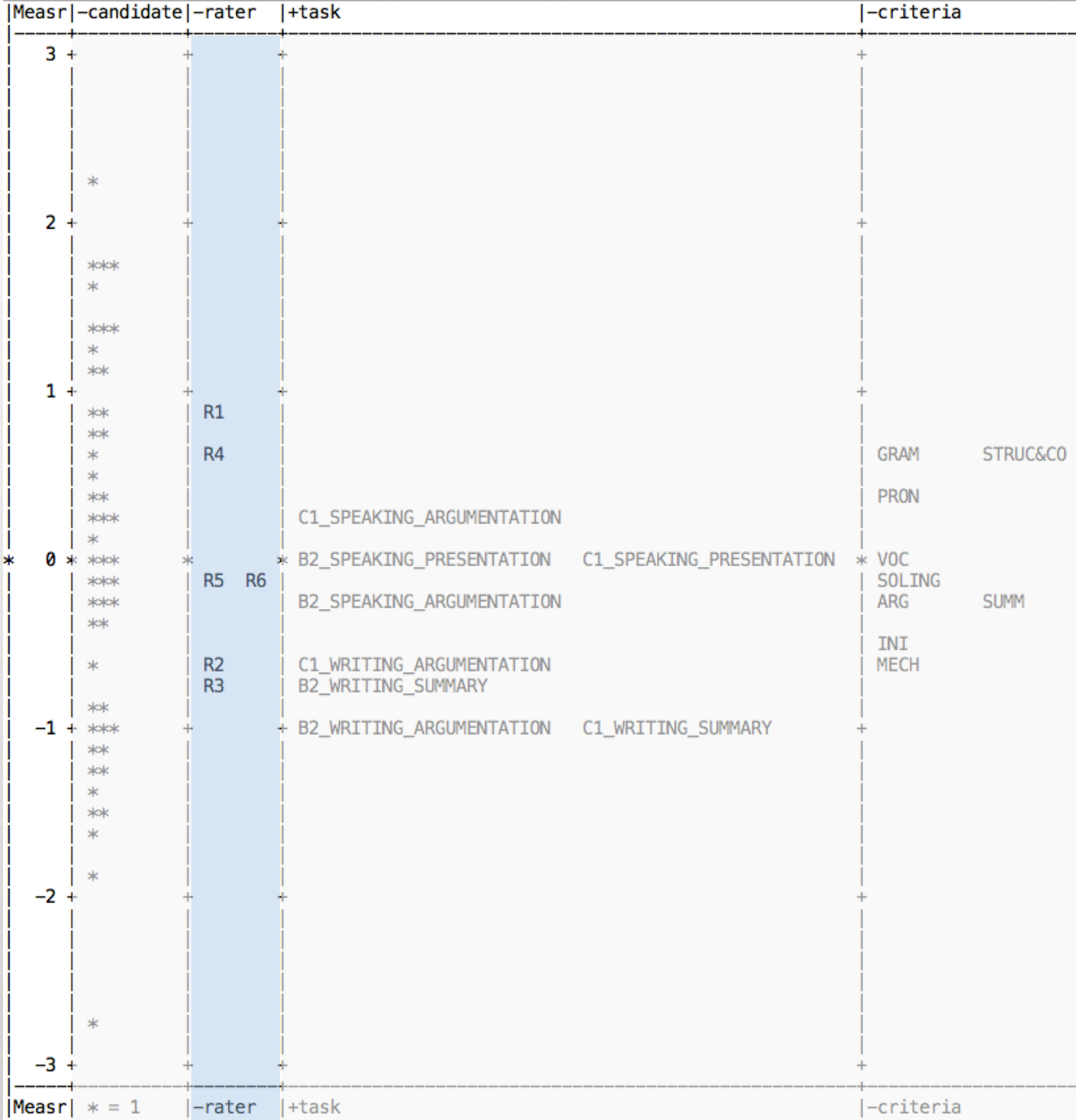
Criteria & score variance

- Levels of difficulty (-.65 – .69)
- Criteria fit the Rasch model (InfitMnSq = .74 – 1.51)
- But some criteria account for more variance than others (argumentation, vocabulary, grammar)

RQ2 Does an empirical co-construction process with novice raters help to stimulate a shared interpretation of CEFR-based criteria?

Qualitative study:

Do raters use the scale in a different way?



Use of rating scale

Sticking to the scale

R2 *Especially when rating oral performances, I had trouble focusing sometimes. I experienced something like an attention problem. After a while you go like "oh yeah, vocabulary – what was that about?"*

I *So do you look at the rating scale again?*

R2 *You keep it at hand of course*

I *Do you use the model differently after the sixtieth time?*

R2 *Gosh*

I *You just said you knew them by heart?*

R2 *Yeah, I kind of do*

R3 *it was more of a wave-like pattern*

R2 *Attention had a lot to do with it*

R1 *Focus, yeah*

Use of rating scale

Sticking to the scale

Scoring band width

R5 *At times I also felt that some performances were really very good and others were just good, but they still got the same level.*

Interpretation of criteria

Sticking to the scale

Scoring band width

Multifaceted criteria

R3 *Grammar just contains so much. There are so many different aspects to consider.*

Interpretation of criteria

Sticking to the scale

Scoring band width

Multifaceted criteria

Rater focus

R2 *For me, layout was important when deciding on "Mechanics". If the punctuation wasn't ok and there was no layout, I'd often assign "C".*

R4 *I really didn't take that into account.*

R5 *I didn't either, but it did bother me sometimes.*

Interpretation of criteria

- Sticking to the scale
- Scoring band width
- Multifaceted criteria
- Rater focus
- Specificity (or the lack of it)

R4: *The individual level descriptors don't always cover all grammatical mistakes, so this can be confusing.*

Interpretation of criteria

Sticking to the scale

Scoring band width

Multifaceted criteria

Rater focus

Specificity (or the lack of it)

Reinterpretation

R4 For me D means that there's an error in every sentence

D "The performance relies mainly on basic syntactic patterns (such as the main clause word order) which may contain mistakes that obscure the meaning of the sentence".

IV Conclusions

RQ1: Reliability

Is it possible to develop a reliable DAP rating scale that is based on CEFR descriptors?

RQ1: Reliability

Is it possible to develop a reliable DAP rating scale that is based on CEFR descriptors?

Quantitatively, yes.

RQ1: Reliability

Is it possible to develop a reliable DAP rating scale that is based on CEFR descriptors?

Qualitatively:

- Some criteria are more important than others
- Descriptor reinterpretation
- Problems stemming from vagueness
- Using scale to fit intuitive judgment
- Perceived unreliability does not translate into operational unreliability (cf. Eckes 2008)

(Weigle, 2002; Lumley, 2002; Cumming et al., 2002; Lumley 2005; Barkaoui, 2010; Eckes, 2008)

RQ2: Co-construction

Does co-construction with raters help to stimulate a shared interpretation of CEFR-based criteria?

RQ2: Co-construction

Does co-construction with raters help to stimulate a shared interpretation of CEFR-based criteria?

Long-term: No

- What is clear to one rater is not self-evident to his/her peer

Short-term: Not really

- Co-construction does not eliminate inherent weaknesses
- Co-construction can act as a catalyst for rater involvement in rater training

Meaning our of vagueness?

- Rating scale co-construction does not eliminate inherent CEFR weaknesses.
- Rating scale co-construction creates rater involvement, but it is a rather impractical way of going about rater standardization.
- The CEFR is an attractive but unstable rating scale foundation.
- Good use of the CEFR demands adaptation, but what is the effect of uncoordinated decentralised adaptations on a gold standard?

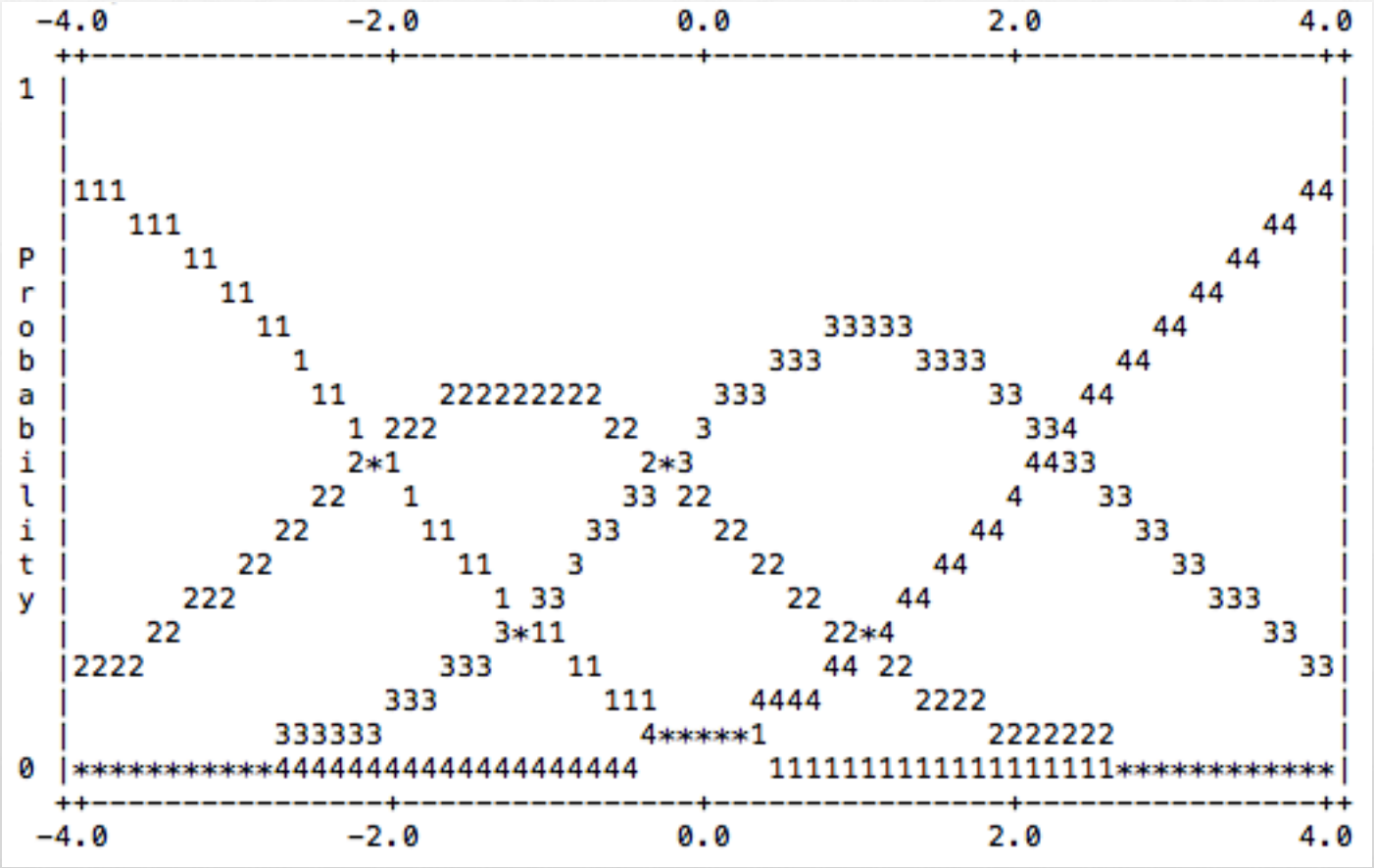
KU LEUVEN

Centrum voor Taal en Onderwijs

Thank you

bart.deygers@arts.kuleuven.be
koen.vangorp@arts.kuleuven.be

Discriminatory potential



Criteria & score variance

	Initial Eigenvalues		
	Tot	% of variance	Cumulative %
Written Argumentation			
Argumentation	1.65	33.06	33.06
Vocabulary	1.21	24.30	57.36
Grammar	.91	18.12	75.48
Written Summary			
Vocabulary	1.71	34.11	34.11
Grammar	1.19	23.86	57.97
Summarizing	.82	16.50	74.47
Spoken Argumentation			
Argumentation	1.64	27.40	27.40
Vocabulary	1.24	20.60	48.01
Grammar	1.01	16.76	64.77
Presentation			
Vocabulary	1.49	24.79	24.79
Grammar	1.12	18.77	43.57
Structure & cohesion	1.02	17.04	60.61